文章编号：1003-0077(2020)08-0061-09

用于社交媒体的中文命名实体识别

李 源，马 磊，邵党国，袁梅宇，张名芳

（昆明理工大学 信息工程与自动化学院，云南 昆明 650504）

摘 要：社交领域的中文命名实体识别(NER)是自然语言处理(NLP)中一项重要的基础任务。目前基于词粒度信息或者外部知识的中文命名实体识别方法，都会受到中文分词(CWS)和溢出词(OOV)等问题的影响。因此，该文提出了一种基于字符的使用位置编码和多种注意力的对抗学习模型。联合使用位置编码和多头注意力能够更好地捕获字序间的依赖关系，而使用空间注意力的判别器则能改善对外部知识的提取效果。该文模型分别在 Weibo2015 数据集和 Weibo2017 数据集上进行了实验，实验结果中的 F_1 值分别为 56.79%和 60.62%。与多个基线模型相比，该文提出的模型性能更优。

关键词：位置编码；多种注意力机制；对抗学习；中文命名实体识别

中图分类号：TP391　　**文献标识码**：A

Chinese Named Entity Recognition for Social Media

LI Yuan, MA Lei, SHAO Dangguo, YUAN Meiyu, ZHANG Mingfang

(Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650504, China)

Abstract: Chinese named entity recognition (NER) in social media is a challenging task. Existing methods based on word-level information or external knowledge are affected by Chinese word segmentation (CWS) and Out-of-Vocabulary (OOV). This paper proposes an adversarial learning model based on character using positional encoding and multi-attention. The combination of positional encoding and self-attention can better capture the dependence of character sequences, while the use of spatial attention discriminator can improve the extraction effect of external knowledge. The experimental results show that the proposed approach achieves 56.79% and 60.62% in *F*-score, respectively, on the datasets in Weibo2015 and Weibo2017.

Keywords: positional encoding; multi-attention; adversarial learning; named entity recognition

0 引言

在网络发展的浪潮之中，微博已成为中国最大的社交媒体平台之一。与此同时，微博也面临低俗炒作、内容暴露等问题。依靠人工处理或者关键词拦截的方法并不能解决大规模、非结构化的网络低俗文本实体识别的问题，因此自动识别非结构化文本中的命名实体就显得十分必要。

传统命名实体识别[1]方法包括基于规则的方法[2]、基于词典的方法[3]和机器学习的方法。例如，基于条件随机场(conditional random field, CRF)[4]的方法。目前，为了能让计算机理解非结构化文本，深度学习已成为处理命名实体识别的最新方法。例如，Huang 等[5]提出使用双向长短时记忆网络(bidirectional long short-term memory, Bi-LSTM)提取单词上下文信息并进行句子标记的 Bi-LSTM-CRF 模型。Peng 和 Dredze[6]则在联合 Bi-LSTM-CRF 进行分词的基础上，进行了联合命名实体识别模型的构建。

同英文相比，中文具有天然的分词劣势。错误的中文分词和存在的溢出词会对下游的任务造成错误的引导[7]。除此之外，与其他自然语言处理的任务相比，命名实体识别任务还存在可用标记语料规

收稿日期：2019-12-02　**定稿日期**：2019-12-31

基金项目：国家自然科学基金(61702069, 61741112)

模较小的问题。为了在中文文本上取得较好的实体识别效果，Zhang 和 Yang[8]从不需要分词的角度出发，在使用包含大量外部知识的词典信息的基础上，提出利用门循环单元自动选择最相关的字词粒度信息的 lattice LSTM 模型，在中文文本上取得较好的识别效果。Zhu 和 Wang[9]在不使用任何外部知识的基础上，提出由带有局部注意输入字符语序的卷积神经网络(convolutional neural networks，CNN)和带有全局自注意力的门控循环网络(gated recurrent unit，GRU)组成的卷积注意力网络(convolutional attention network，CAN)模型。Cao 等[10]提出了一种使用基于字符粒度的，借助 Bi-LSTM 提取共用特征进行命名实体识别和中文分词两个任务的多任务网络框架。从对多任务识别的角度出发，增加了可用的命名实体任务的标记语料规模。

但是，以上为中文文本设计的解决方案依然存在一些问题。其中一个问题是与其他的中文文本相比，微博的中文文本更加不规则，同时也包含了大量溢出词。因此，以 Zhang 和 Yang[8]为代表的方法会受到外部词典的限制。而另一个问题是以上方法没有综合考虑输入字符的语序依赖和外部知识的共同作用。例如，Zhang 和 Yang[8]和 Cao 等[10]均没有考虑输入字符语序依赖对识别的影响。而 Zhu 和 Wang[9]的方案存在的问题是不能从外部知识中提取信息。

综上，社交媒体的中文命名实体识别任务具有分词劣势、多溢出词、对字符语序间依赖以及对外部知识的共同作用探索不足等问题，因此本文提出了一种基于字符使用位置编码和多种注意力的对抗学习模型。

本文的贡献总结如下：

(1) 在对抗训练的过程中，提出了一种新的应用空间注意力机制的判别器模型。同时，通过分析消融实验中的结果，验证了本文提出的判别器模型在提升模型性能上的有效性。

(2) 将字符粒度的位置编码和多头注意力机制应用到对抗学习模型中，能够解决错误分词、溢出词和输入字符间依赖的问题。实验结果也表明，联合使用位置编码和多头注意力机制能够有效地提升模型捕获输入字序间依赖的性能。除此之外，本文模型与多个基线模型相比均取得了更好的性能。

1 方法

本节先给出本文模型的整体框架，如图 1 所示，包括三个部分：①信息提取部分，提取不同字符间依赖和上下文特征信息；②序列标注部分，利用在信

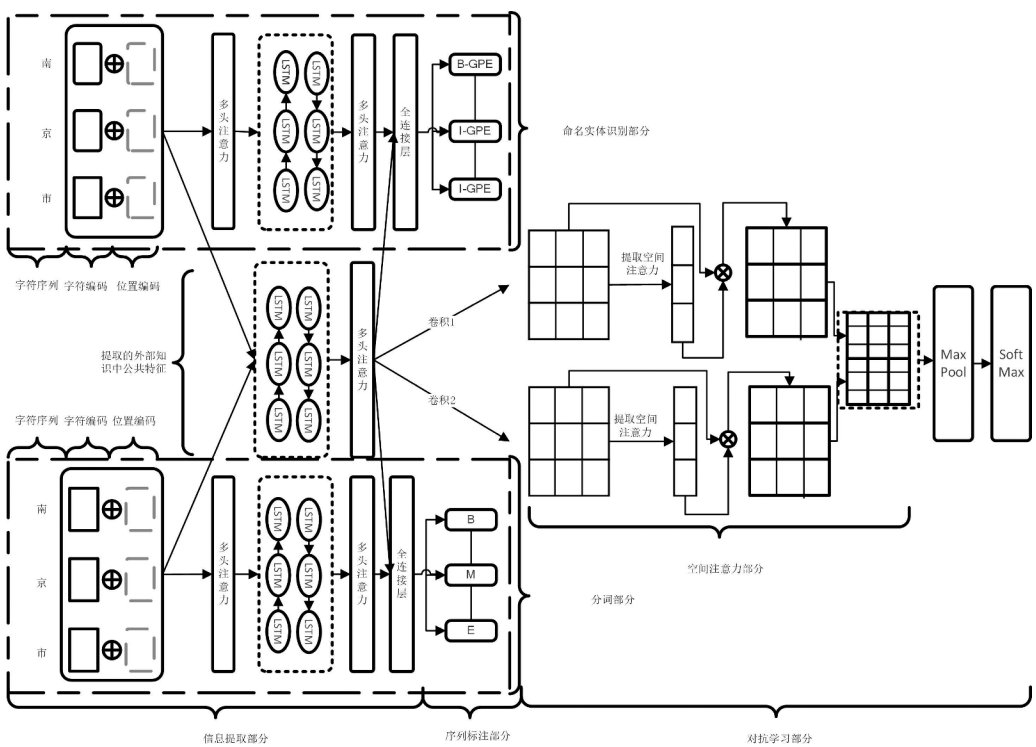

图 1 本文模型网络框架

息提取部分提取的任务特征和通过外部知识捕获的公共特征进行序列标注；③对抗学习部分，模型通过每次交替训练命名实体识别任务和分词任务来提取外部知识中的公共特征，即在实现最大化公共特征的同时，使分词任务和命名实体识别任务的损失达到最小。本节将详细介绍模型的各个组成部分。

1.1　信息提取部分

如何有效捕获非结构化文本的特征信息是命名实体识别任务的核心之一。受到认知理论[11]的启示，并基于人类阅读的实际情况，本文了解到外部知识、对重点信息的把控和阅读顺序能够有效地提高人类对句子的理解程度。因此，本文将从这三个方面建模文本的特征信息。

1.1.1　嵌入层

由于分词任务具有较多的与命名实体任务相同的标记实体边界，而且分词的语料规模又相对可观，所以本文选取分词语料作为外部知识并设计了用于分词任务的字符向量 $\boldsymbol{e}^{\mathrm{CWS}}$ 和用于命名实体识别任务的字符向量 $\boldsymbol{e}^{\mathrm{NER}}$ 分别作为模型的输入向量。其中，$\boldsymbol{e}^{\mathrm{CWS}}$ 中包含能为命名实体识别任务的边界判定提供划分依据的外部知识和一定程度的噪声，本文将在对抗学习部分进行噪声处理。$\boldsymbol{e}^{\mathrm{NER}}$ 则能提供独属于命名实体识别的语义特征。通过对相应的词向量矩阵的查询可得这两类字符的特征表示，如式(1)所示。

$$\boldsymbol{e}^t = \{e_1^t, e_2^t, e_3^t, \cdots, e_n^t\} \tag{1}$$

其中，n 为输入句子的长度，$\mathrm{T}=\{\mathrm{t}\mid \mathrm{t}=1, \mathrm{t}=2\}$ 表示任务类别集合。例如，t=1 表示现在的输入字符向量是分词任务的向量，t=2 则表示现在的输入字符向量是命名实体识别任务的向量。$\boldsymbol{e}^t \in R^{n\times d_1}$ 表示在处理 t 任务时，输入句子对应的字符向量矩阵集合，d_1 表示嵌入维度。

1.1.2　使用位置编码的多头注意力机制

鉴于句子中多个实体之间可能存在较强的依赖关系。受到 Ashish Vaswani 等人工作[12]的启发，本文引入位置编码信息和多头注意力机制来改善模型的效果。多头注意力通过多个表示子空间扩展了模型专注不同位置的能力，但是单独的多头注意力机制缺少对输入单词顺序的理解。因此，本文使用了结合位置编码的多头注意力机制来提取输入相邻字符的依赖关系，计算多头注意力 A(Q,K,V)的公式，如式(2)所示。

$$A(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_1}}\right)\boldsymbol{V} \tag{2}$$

其中，$\boldsymbol{Q}\in R^{n\times d_2}$ 代表查询矩阵，$\boldsymbol{K}\in R^{n\times d_2}$ 代表键矩阵，$\boldsymbol{V}\in R^{n\times d_2}$ 代表值矩阵。而位置编码的公式如式(3)～式(8)所示。

$$\mathrm{PE}(\mathrm{pos}, 2i) = \sin\left(\frac{\mathrm{pos}}{10\ 000^{\frac{2i}{d_1}}}\right) \tag{3}$$

$$\mathrm{PE}(\mathrm{pos}, 2i+1) = \cos\left(\frac{\mathrm{pos}}{10\ 000^{\frac{2i}{d_1}}}\right) \tag{4}$$

$$\boldsymbol{p}_j^t = [\mathrm{PE}(j,0); \mathrm{PE}(j,1); \cdots; \mathrm{PE}(j,d_1)] \tag{5}$$

$$\boldsymbol{p}^t = [p_1^t, p_2^t, p_3^t, \cdots, p_n^t] \tag{6}$$

$$\mathrm{head}_k^t = A[(\boldsymbol{e}^t+\boldsymbol{p}^t)\boldsymbol{w}_k^{<q,t>}, (\boldsymbol{e}^t+\boldsymbol{p}^t)\boldsymbol{w}_k^{<k,t>}, (\boldsymbol{e}^t+\boldsymbol{p}^t)\boldsymbol{w}_k^{<v,t>}] \tag{7}$$

$$H^t = [\mathrm{head}_1^t; \mathrm{head}_2^t; \cdots; \mathrm{head}_h^t]\boldsymbol{w}_0 \tag{8}$$

其中，pos 是当前字符在句子中的位置，i 表示字符嵌入维度的维度位置，$PE(pos,i)$ 代表在第 pos 个位置的字符向量的第 i 个维度的位置编码。$\boldsymbol{p}_j^t \in R^{1\times d_1}$ 代表在处理 t 任务时，第 j 个字符的位置编码。$\boldsymbol{p}^t \in R^{n\times d_1}$ 代表当前输入句子的位置编码矩阵。$\boldsymbol{w}_k^{<q,t>}\in R^{d_1\times d_2}$、$\boldsymbol{w}_k^{<k,t>}\in R^{d_1\times d_2}$、$\boldsymbol{w}_k^{<v,t>}\in R^{d_1\times d_2}$ 表示在处理 t 任务时，第 k 个表示子空间的 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别对应的转换矩阵，h 表示设置的表示子空间的个数，$d_2=\dfrac{d_1}{h}$，$\mathrm{head}_k^t \in R^{n\times d_2}$ 表示对输入句子的第 k 个表示子空间提取的特征。$\boldsymbol{w}_0 \in R^{d_1\times d_1}$ 表示将多个表示子空间提取的不同维度的位置注意力特征映射到同一维度空间的转换矩阵，$\boldsymbol{H}^t \in R^{n\times d_1}$ 表示对输入的句子注意重要信息的结果。

1.1.3　Bi-LSTM 层

因为句子中双向信息有助于序列建模，所以本文采用能捕获文本双向信息的 Bi-LSTM[13] 网络来进行句子上下文特征的提取，如式(9)～式(12)所示。

$$\overrightarrow{h_i^t} = \overrightarrow{\mathrm{LSTM}}(\overrightarrow{a^{<i-1>}}, H_i^t) \tag{9}$$

$$\overleftarrow{h_i^t} = \overleftarrow{LSTM}(\overleftarrow{a^{<i-1>}}, H_i^t) \tag{10}$$

$$h_i^t = [\overrightarrow{h_i^t}, \overleftarrow{h_i^t}] \tag{11}$$

$$h_c^t = [h_1^t, h_2^t, h_3^t, \cdots, h_n^t] \tag{12}$$

其中，$a^{<i>}$ 表示当前记忆单元的隐藏层状态。$\overrightarrow{h_i^t}$ 和 $\overleftarrow{h_i^t}$ 分别表示在处理 t 任务时，在第 i 个字符位置的过去和未来记忆网络的隐藏状态。h_i^t 表示这两个方向隐藏状态的组合。$\boldsymbol{h}_c^t \in R^{n\times d_3}$ 是对在处理 t 任务时，对输入句子上下文信息进行编码的结果，d_3 表示 Bi-LSTM 的输出维度。虽然 Bi-LSTM 能够获取语序信息，并能在短句中较好地捕获实体之间的前后依赖，但是在处理长句时，该机制就很难捕获

实体间的依赖关系。因此在式(12)之后,本文使用了多头注意力机制来捕获高维上下文特征间的依赖关系。其注意力公式如式(2)所示,不再赘述。同理,根据式(9)~式(12),$\boldsymbol{h}_c^s$ 则表示通过 Bi-LSTM 层提取的外部知识中公共特征,对应于图 1 左中部提取的外部知识中的公共特征部分。模型通过对抗训练来保证在 $\boldsymbol{h}_c^s$ 存储的特征为公共特征,在消融模型中的实验结果也验证了这一观点。

1.2 序列标注部分

命名实体识别和分词任务都是序列标注任务,相邻的标签之间有很强的约束关系。因此,本文采用了 CRF[14] 作为解码层。简单来说,CRF 是由标签概率矩阵 $\boldsymbol{E}\in\boldsymbol{R}^{n\times \text{tags}}$ 和转移概率矩阵 $\boldsymbol{T}\in R^{\text{tags}\times\text{tags}}$ 构成,其中 n 为句子中字符的个数,tags 为标签的个数。本文通过维特比算法来最小化 t 任务的损失函数,并找出最高概率的句子序列,如式(13)~式(16)所示。

$$\boldsymbol{E}=\sigma([\boldsymbol{h}_c^t;\boldsymbol{h}_c^s]\boldsymbol{w}_t+\boldsymbol{b}_t) \tag{13}$$

$$s(\boldsymbol{S},y)=\sum_{i=1}^{n}\boldsymbol{E}_{i,y_i}+\sum_{i=1}^{n-1}\boldsymbol{T}_{y_i,y_{i-1}} \tag{14}$$

$$p(y\mid\boldsymbol{S})=\frac{e^{s(\boldsymbol{S},y)}}{\sum_{y'\in Y}e^{s(\boldsymbol{S},y')}} \tag{15}$$

$$\text{loss}^t=-\log p(\hat{y}^t\mid S) \tag{16}$$

其中,$\boldsymbol{w}_t\in R^{d_4\times\text{tags}}$、$\boldsymbol{b}_t\in R^{n\times\text{tags}}$ 是全连接层的参数 $d_4=2\times d_3$。Y 是所有输出序列的集合。$\hat{y}^t$ 是在处理 t 任务时的实际标签。

1.3 对抗学习部分

虽然分词语料中的外部知识能够为命名实体识别任务的实体边界识别提供共性信息,但是这些外部知识也会引入一定程度的噪声。例如,在命名实体识别任务中被标记为一个实体的"休斯顿机场"在分词任务中却被标记划分为"休斯顿"和"机场"。对于命名实体识别任务来说"休斯顿"和"机场"就是噪声。受到 Cao 等[10] 使用最大池化和 Softmax 组成的判别器过滤独属于分词任务特有信息的启发,本文提出了一种新的基于空间注意力机制的判别器模型。实验结果表明本文的判别器模型能够更加有效地提取公共信息,从而提高模型的性能。

1.3.1 空间注意力机制

原本的空间注意力[15] 是指对图像进行特征提取时,找到更加重要的像素特征位置的机制。因为文本并没有图像中的频道,所以在这里进行一定程度的修改。为了更高效地提取有效信息,本文在使用两种不同的卷积核进行不同维度特征提取的基础上,使用最大池化操作和平均池化操作分别保留文本中局部最重要的特征和平均特征。然后,再对池化的结果进行卷积操作,并使用卷积神经网络来模拟这些特征的重要程度,从而得出对每个输入字的注意力大小。注意力系数的计算如式(17)~式(18)所示。

$$F_i^s=f^{i\times d_3}(\boldsymbol{h}_c^s) \tag{17}$$

$$M^s(F_i^s)=\sigma(f^{7\times7}([\text{AvgPool}(F_i^s);\text{MaxPool}(F_i^s)])) \tag{18}$$

其中,$i\in\{3,5\}$。$f^{i\times d_3}$ 表示使用 $i\times d_3$ 的卷积核对输入的文本向量进行卷积操作。F_i^s 表示卷积的结果。AvgPool、MaxPool 分别表示平均池化操作和最大池化操作。最后再按照提取特征的重要程度分别对这两种卷积的结果筛选出其对应的前 n 个最大特征,并将它们组合起来,如式(19)所示。

$$\begin{aligned}\text{sa}^s=[&\text{MaxFeature}(M^s(F_3^s)\times F_3^s);\\&\text{MaxFeature}(M^s(F_5^s)\times F_5^s)]\end{aligned} \tag{19}$$

其中,MaxFeature 表示按照重要程度筛选出前 n 个特征。$\text{sa}^s\in R^{n\times d_3}$ 代表使用两个不同的卷积操作提取的有效特征。空间注意力结构如图 2 所示。

1.3.2 基于空间注意力的判别器

为了减少外部知识中的噪声对命名实体识别任务的影响,本文使用了对抗学习[16] 的思想,即通过使判别器不能识别当前提取特征的任务归属来实现对共性特征的最大提取,如式(20)~式(23)所示。

$$s^s=\text{MaxPool}(sa^s) \tag{20}$$

$$D(s^s,\theta_s)=\text{softmax}(\boldsymbol{W}_s s^s+\boldsymbol{b}_s) \tag{21}$$

$$\text{loss}_{\text{Adv}}=\max_{\theta_e}\left(\min_{\theta_s}\sum_{t=1}^{T}\log D(E_e(H^t))\right) \tag{22}$$

$$\text{loss}=\sum_{t=1}^{T}(t-1)\times(\text{loss}^{t-1})-\lambda\,\text{loss}_{\text{Adv}} \tag{23}$$

其中,T 表示任务的个数,$\boldsymbol{W}_s\in\text{R}^{T\times d_3}$ 和 $\boldsymbol{b}_s\in R^T$ 是全连接层的参数,$D(s^s,\theta_s)$ 表示判别器的损失,θ_s 表示在判别器识别中的训练参数,$E_e(H^t)$ 表示提取的公共特征信息,θ_e 表示在提取公共特征中的训练参数,λ 是用来权衡学习效果的超参数。式(23)则表示一个输入句子的对抗损失函数。通过在训练集上的训练,模型最后会达到一个公共特征误差最小而判别器的损失最大的平衡点。在此平衡点处,模型达到 $E_e(H^t)$ 存储的外部知识的公共信息最大,而噪声信息最小的状态。

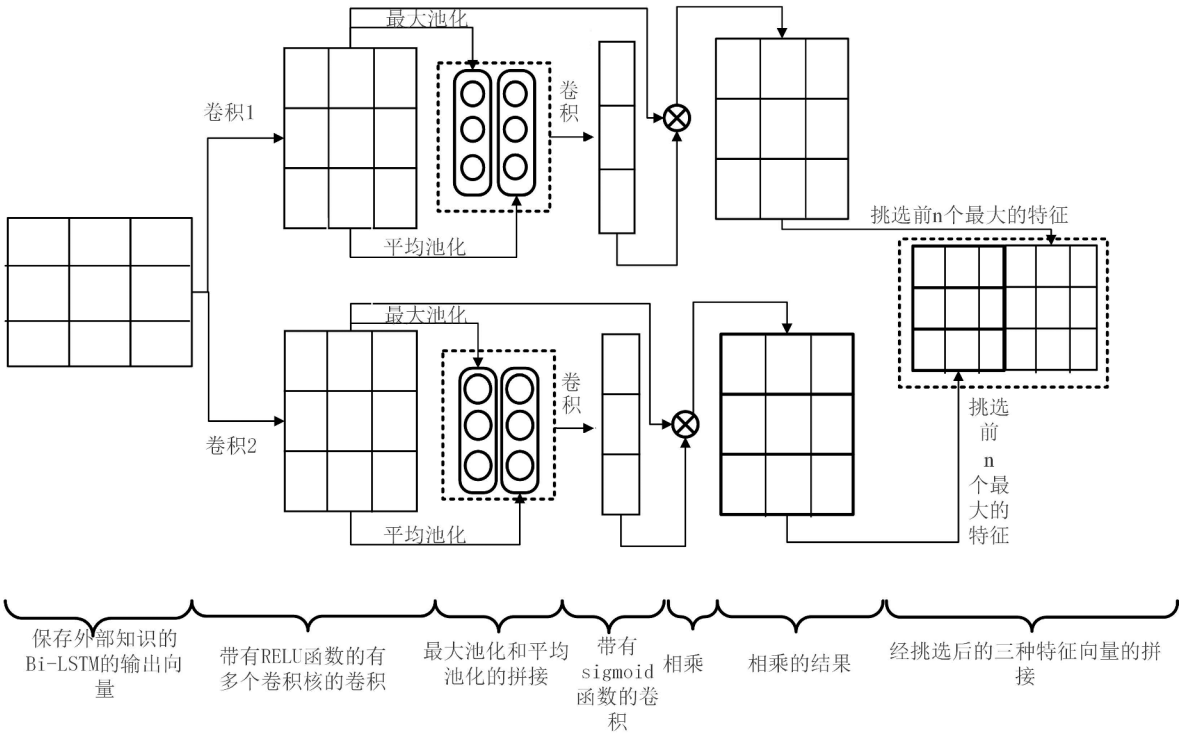

图 2　空间注意力结构图

2　实验与分析

2.1　实验数据

本文选取 Weibo2015 和 Weibo2017[①] 数据集作为实验数据集，并使用 SIGHAN 2005[②] 分词数据集作为外部知识数据集。Weibo2015 和 Weibo2017 是包含了人名、地名、组织名、地缘政治实体四种实体的公开数据集。数据详情如表 1 所示。

表 1　实验数据统计

数据集	训练集	验证集	测试集	实体类型数目
Weibo2015	1 350	270	270	4
Weibo2017	1 350	270	270	4
SIGHAN 2005	86 924	—	3 985	—

2.2　实验参数设置

本文使用文献[10]训练的词向量，并将实验的字符向量维度设为 100。如果在使用的词向量中未找到当前训练的单词，则由[−1,1]均匀分布随机初始化的生成向量来表示。对于其他超参数，则取值如表 2 所示。

表 2　模型的超参数设置

超参数	数值
Mini-batch 大小	20
多头自注意力层数	2
Bi-LSTM 层数	1
空间注意力层数	1
λ	0.6
Learning rate	0.001
optimization	Adam

2.3　评估方法

本文所采用的评价指标是整体的准确率(P)、召回率(R)和 F_1-score(F_1)，其中每一个类别的指标的计算如式(24)～式(26)所示。

$$P = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{24}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{25}$$

$$F_1 = \frac{2PR}{P+R} \tag{26}$$

其中，TP 是真阳的数目，FP 是假阳的数目，FN

① https://github.com/hltcoe/golden-horse/tree/master/data

② http://sighan.cs.uchicago.edu/bakeoff2005/

是假阴的数目。

2.4 对比模型

为了验证本文所提模型的有效性，选择以下与本文相关的具有代表性的系统作为基线模型。

(1) Peng 和 Dredze[4] 提出联合字符位置词嵌入信息的 CRF 联合模型。

(2) Huang 等[5] 通过 Bi-LSTM 提取语序特征，并利用 CRF 进行标签预测。

(3) Peng 和 Dredze[6] 在使用 Chen 等[17] 的基于 Bi-LSTM-CRF 的中文分词模型的基础上，增强了 Peng 和 Dredze[4] 的模型性能，获得了更好的识别性能。

(4) He 和 Sun[18] 将跨领域学习和半监督学习联合起来进行中文社交领域的命名实体识别。

(5) Cao 等[10] 利用对抗学习进行命名实体识别。该模型通过在字符粒度上，提取分词任务和命名实体识别任务的公共词边界信息来获取更多共性特征信息。

(6) Zhang 和 Yang[8] 通过词典匹配单词来查找相关词向量。他们在引入了单词粒度信息的同时，避免了分词的错误，最后再将匹配的单词和字符送入 lattice LSTM 以提取最相关的字和词粒度的信息。

(7) Zhu 和 Wang[9] 为了避免错误分词和溢出词引入的噪声信息，首次将卷积神经网络和局部注意力机制联合起来，增强模型捕获字符序列的能力。

2.5 实验结果与分析

对比模型在 Weibo2015 上的实验性能如表 3 所示。

表 3 对比模型在 Weibo2015 上的实验性能

模型	P	R	F_1
Peng 和 Dredze[4]	57.98	35.57	44.09
Peng 和 Dredze[6]	**63.33**	39.18	48.41
Huang 等[5]	58.99	44.93	51.01
Cao 等[10]	55.72	50.68	53.08
本文模型	60.52	**53.48**	**56.79**

通过观察表 3 可以得到以下结论：

整体上，Peng 和 Dredze[4] 的性能低于其他使用深度学习的方法，导致这种现象的原因可能是基于深度学习的方法比起只使用 CRF 的方法能更加容易捕捉到字或词内部的高维深度语义特征。另外，与 Peng 和 Dredze[6] 和 Huang 等[5] 相比，Cao 等[10] 能够取得更高的 F_1 值。这可能是因为这两个模型虽然使用了字和词粒度的信息，但是由于可能存在的错误分词和溢出词问题，而致使模型的性能下降。与之相比，Cao 等[10] 在引入了字粒度分词信息的基础上，使用最大池化来提取公共特征，最终提高命名实体识别的性能。本文模型则在考虑输入字符粒度位置信息和外部字符粒度共用知识信息的同时，使用空间注意力机制处理了判别器中最大池化可能造成的词序信息丢失问题[19]，从而取得了比其他模型高的 F_1 值。

具体地，作为深度学习的代表方法 Peng 和 Dredze[6] 的方法识别性能最低。而 Cao 等[10] 在 baseline 中取得了 53.08% 的最高 F_1 值，这表明字粒度的外部知识特征对于识别的有效性。然而，该模型没有对词序信息丢失问题进行处理。因此本文模型与之相比在 R 值上获得 2.8% 提升，这可能是位置编码和多种注意力机制的共同作用效率。

对比模型在 Weibo2017 上的实验性能如表 4 所示。

表 4 对比模型在 Weibo2017 上的实验性能

模型	P	R	F_1
Peng 和 Dredze[4]	**74.78**	39.81	51.96
Peng 和 Dredze[6]	66.67	47.22	55.28
He 和 Sun[18]	61.68	48.82	54.50
Cao 等[10]	59.51	50.00	54.34
Zhang 和 Yang[8]	—①	—	53.04
Zhu 和 Wang[9]	—	—	55.38
本文模型	67.29	**55.15**	**60.62**

综合表 3 和表 4 的结果可以看出本文模型的 R 值和 F_1 值在这两个数据集上均为最高。我们也可以得到与表 3 中相似的结论，但这两张表之间也存在部分模型性能排序上的不同。具体表现为，表 4 中 Cao 等[10]、He 和 Sun[18] 和 Zhang 和 Yang[8] 的 F_1 值分别要比 Peng 和 Dredze[6] 的 F_1 值低 0.94%、0.78% 和 2.24%。究其原因，可能是因为 Weibo2017 的语料比 Weibo2015 更规范，所以使用 Weibo2017 作为语料的模型更不容易受到错误分词和溢出词的影响。因此，Peng 和 Dredze[6] 使用的

① 未获取到 Zhang 和 Yang 和 Zhu 和 Wang 成果

Bi-LSTM 会从训练语料中提取更少的分词错误，CRF 也能学到更加有效的语法规则，从而使得该模型的性能更好。有必要指出的是 Zhang 和 Yang[8] 的模型性能要远低于其他模型。这可能是受到 Weibo 语料中大量新词和不规范缩写词的影响，使得该模型不能从预定义的外部字典中匹配到有效单词，从而使得模型损失大量词粒度的信息，最终导致模型性能下降。

同样值得注意的是，同表 4 中 baseline 间 F_1 值最高的 Zhu 和 Wang[9] 和使用外部知识的 Peng 和 Dredze[6] 的模型相比，本文模型在性能上分别取得了 5.24%和 5.34%的提升。这可能是因为同不使用任何外部知识的 Zhu 和 Wang[9] 相比，本文模型使用了更加有效的外部共用知识提取手段，而这种外部共用知识特征恰好能弥补单独字符特征缺失的高维语义关系。而 Peng 和 Dredze[6] 虽然也使用了相同的外部语料，但是他们没有将外部知识中的噪声信息有效去除，因此造成模型整体效果不好。与该两组模型的比较也验证了本文设计的提取外部共用知识方法的有效性。

2.6　消融分析

为了验证本文提出方法中位置编码、多头注意力机制以及在对抗学习中使用的判别器的有效性，本文设计以下模型并在 Weibo 数据集上进行了实验与分析。

（1）**Bi-LSTM-CRF**　使用基于字符的 Bi-LSTM 提取过去和未来上下文信息。

（2）**Bi-LSTM-CRF-adv**　在 Bi-LSTM-CRF 的基础上，只使用对抗学习提取命名实体识别任务和分词任务的共用特征。

（3）**Bi-LSTM-CRF-adv ＋ self-attention**　在 Bi-LSTM-CRF-adv 的基础上，在模型嵌入层后使用多头自注意力机制。

（4）**Bi-LSTM-CRF-adv ＋ position**　在 Bi-LSTM-CRF-adv 的基础上，在模型嵌入层使用位置编码。

（5）**Bi-LSTM-CRF-adv ＋position＋ self-attention**　在 Bi-LSTM-CRF-adv ＋position 的基础上，在模型嵌入层后使用多头自注意力机制。

（6）**Bi-LSTM-CRF-adv ＋ Spatial Attention**　在 Bi-LSTM-CRF-adv 的基础上，使用本文提出的包含空间注意力判别器的模型进行对抗学习。

表 5 给出 Bi-LSTM-CRF 和 Bi-LSTM-CRF-adv 的实验结果，并将其作为基线方法。其中 Bi-LSTM-CRF的三项指标在两个数据集上均低于 Bi-LSTM-CRF-adv，这显示了通过对抗学习提取外部共用知识的有效性。

表 5　消融模型的实验性能

模型	Weibo2015			Weibo2017		
	P	R	F_1	P	R	F_1
Bi-LSTM-CRF	58.99	44.93	51.01	65.87	39.71	49.55
Bi-LSTM-CRF-adv	59.23	49.48	53.92	**69.62**	48.45	57.14
Bi-LSTM-CRF-adv ＋ self-attention	56.44	53.48	54.92	67.58	50.51	57.81
Bi-LSTM-CRF-adv ＋position	52.90	52.90	52.90	62.26	51.03	56.09
Bi-LSTM-CRF-adv ＋position＋ self-attention	54.65	**54.65**	54.65	59.25	**57.77**	58.48
Bi-LSTM-CRF-adv ＋ Spatial Attention	57.05	51.74	54.26	65.60	53.09	58.68
本文模型	**60.52**	53.48	**56.79**	67.29	55.15	**60.62**

同 Bi-LSTM-CRF-adv 相比，单独使用位置编码的模型、单独使用多头自注意力机制的模型，以及联合使用位置编码和多头自注意力机制的模型均能大幅度提高这些模型的 R 值，尤其是联合使用位置编码和多头自注意力机制的模型对 R 值的提升效果最好。但 Bi-LSTM-CRF-adv ＋position 的整体性能值却不如 Bi-LSTM-CRF-adv。这可能是因为单独使用位置编码会增加模型复杂度，使得模型难以从只使用位置编码的模型中提取更为有效的词序信息。除此之外，同 Bi-LSTM-CRF-adv 相比，单独使用多头自注意力机制虽然能提升模型性能，但是却缺失对嵌入层输入字序的信息。而联合使用这两者能够弥补它们的不足之处，即位置编码为多头自注意力提供了词嵌入的语序信息，而这一信息又使

得多头自注意力能够更加有效地捕获词嵌入的有效特征。

同时，有必要指出的是 Bi-LSTM-CRF-adv ＋ position ＋ self-attention 的 F_1 值与 Bi-LSTM-CRF-adv ＋ self-attention 的 F_1 值在两个数据集上的大小关系不同。这可能是因为 Weibo2017 比 Weibo2015 的数据更加规范，而 Bi-LSTM-CRF-adv ＋position＋ self-attention 在规范的数据集上又更容易捕获语义依赖关系所致。以上结果表明，综合使用位置编码和多头自注意力机制的有效性，其案例如表 6 所示。

表 6 案例 1

①	根	据	发	改	委	发	布	通	知
adv	O	O	O	O	O	O	O	O	O
ps	O	O	B-ORG	I-ORG	I-ORG	O	O	O	O
Gold	O	O	B-ORG	I-ORG	I-ORG	O	O	O	O

同 Bi-LSTM-CRF-adv 相比，Bi-LSTM-CRF-adv＋ Spatial Attention 的 R 值和 F_1 值在两个数据集上均有一定程度的提升。这表明使用包含空间注意力的判别器能更加有效地捕获外部共用知识。同 Bi-LSTM-CRF-adv ＋position＋ self-attention 相比，使用了空间注意力的模型的 P 值更高。这一结果表明使用包含空间注意力的判别器能够捕获一些 Bi-LSTM-CRF-adv ＋position＋ self-attention 所缺失的但有助于最终识别的特征信息。因此，Bi-LSTM-CRF-adv＋ Spatial Attention 的模型能够较好地提高 P 值。其案例如表 7 所示。

表 7 案例 2

	宋	同	志	到	汉	口	站
adv	B-PER	I-PER	I-PER	O	B-LOC	I-LOC	I-LOC
sa	O	O	O	O	B-LOC	I-LOC	I-LOC
Gold	O	O	O	O	B-LOC	I-LOC	I-LOC

最后，以上变体模型的 $F1$ 值均低于本文模型。这可能是外部知识、位置编码和多种注意力综合作用的结果。

3 结论

针对中文命名实体识别问题，本文提出了一种基于字符的使用位置编码和多种注意力的对抗学习模型。本文模型通过联合位置编码和多头自注意力来捕获字序间的依赖关系，并为之后的上下文信息提取提供了更有效的特征信息；同时由于判别器的性能是提取公共特征的决定因素，所以本文使用空间注意力机制来提高判别器的提取效果，最终与不同模型的对比结果也验证了本文提出的判别器模型和整体模型的优越性。

参考文献

[1] Chinchor N. MUC-6 named entity task definition[C]//Proceedings of the 6th Conference on Message Understanding, 1995.

[2] Fukuda K, Tsunoda T, Tamura A, et al. Toward information extraction: identifying protein names from biological papers[C]//Proceedings of the Pacific Symposium on Biocomputing. 1998.

[3] Rindflesch T C, Tanabe L, Weinstein J N, et al. EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature[C]//Proceedings of the Pacific Symposium on Biocomputing. 2000.

[4] Nanyun Peng, Mark Dredze. Named entity recognition for chinese social media with jointly trained embeddings[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 548-554.

[5] Zhiheng Huang, Wei Xu, Kai Yu. Bidirectional LSTM-CRF models for sequence tagging [J]. arXiv preprint arXiv: 1508.01991, 2015.

[6] Nanyun Peng, Mark Dredze. Improving named entity recognition for chinese social media with word segmentation representation learning[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016:149.

[7] Xiaoya Li, Yuxian Meng, Xiaofei Sun, et al. Is Word Segmentation Necessary for Deep Learning of Chinese Representations? [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL, 2019:3242-3252.

[8] Yue Zhang, Jie Yang. Chinese NER using lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL,2018, 1:1554-1564.

[9] Yuying Zhu, Guoxin Wang. CAN-NER: Convolution-

① 本文使用 adv、ps、sa 分别代表 Bi-LSTM-CRF-adv、Bi-LSTM-CRF-adv ＋position＋ self-attention、Bi-LSTM-CRF-adv ＋Spatial Attention 模型的预测结果。而 Gold 代表真标签。

al Attention Network for Chinese Named Entity Recognition[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA ,2019 ,1:3384-3393.

[10] Pengfei Cao, Yubo Chen, Kang Liu, et al. Adversarial Transfer Learning for Chinese Named Entity Recognition with Self-Attention Mechanism [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 2018: 182-192.

[11] Wren S. The Cognitive Foundations of Learning to Read : A Framework, ED448420 [R]. Washington, DC:Office of Educational Research and Improvement, 2001.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is All you Need[J]. arXiv preprint arXiv: 1706.03762, 2017.

[13] Mike Schuster, Kuldip K. Paliwal. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11):2673-2681.

[14] John Lafferty, Andrew McCallum, Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data [C]//Proceedings of the 18th International Conference on Machine Learning, Williams College, Williamstown, MA, USA, 2001: 282-289.

[15] Sanghyun Woo, Jongchan Park, Joon Young Lee. CBAM: Convolutional Block Attention Module[C]//Proceedings of the European Conference on Computer Vision, 2018: 3-19.

[16] Goodfellow I J, Pouget Abadie J, Mirza M, et al. Generative Adversarial Nets[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. 2014.

[17] Xinchi Chen, XipengQiu, Chenxi Zhu, et al. Long short-term memory neural networks for chinese word segmentation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2015: 1197-1206.

[18] Hangfeng He, Xu Sun. A unified model for cross-domain and semi-supervised named entity recognition in chinese social media[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2017: 3216-3222.

[19] Hu Xu, Bing Liu, Lei Shu, et al. Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 2018,2:592-598.

李源(1995—),硕士研究生,主要研究领域为数据挖掘、自然语言处理。
E-mail: yuanl1995@qq.com

邵党国(1979—),博士,副教授,主要研究领域为数据挖掘、文本处理。
E-mail: huntersdg@163.com

马磊(1978—),通信作者,工程师,主要研究领域为数据挖掘、人工智能。
E-mail: roy_murray@qq.com